DOCUMENT RESUME

ED 243 899 TM 830 613

AUTHOR Lindstrom, Berner
TITLE The Rasch Model as a Criterion: Applying the Rasch Model to the Analysis of Test Heterogeneity.
SPONS AGENCY National Swedish Board of Education, Stockholm.
PUB DATE Apr 83
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Latent Trait Theory; *Mathematical Models; Multidimensional Scaling; *Psychometrics; Test Construction; Tests; *Test Theory
IDENTIFIERS Item Parameters; Person Parameters; *Rasch Model; *Test Homogeneity; Unidimensionality (Tests)

ABSTRACT

The aim of this paper is to explore the Rasch model as a criterion of test homogeneity. Two empirical studies are presented to demonstrate this usage. From these studies it is argued that statistical tests of item characteristic curve (ICC) slopes are not sufficient in testing for heterogeneity. Tests of equality of ICC's over groups of subject and, above all, tests of equality of person characteristic curve's over groups of items are the proper tests to use. It is also stressed that the application of these kinds of test puts high demands on a theoretical foundation. Finally, it is maintained that the problem of test heterogeneity is essentially a theoretical question. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED243899

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. Lindström

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THE RASCH MODEL AS A CRITERION

Applying the Rasch model to the analysis of test heterogeneity

Berner Lindström
University of Göteborg

Paper to be presented at the Annual Meeting of the American Educational Research Association, Montreal, April 11-15, 1983. Requests for copies should be sent to Berner Lindström, Department of Education, University of Göteborg, Box 1010, S-431 26 Mölndal, Sweden.

TM 830 613

The work presented in this paper has been supported financially by the Swedish National Board of Education. I also want to express my gratitude to Professor Hans Spada for his willingness to give access to his data and for valuable comments on my reanalyses.

ABSTRACT

The aim of this paper is to explore the Rasch model as a criterion of test homogeneity. Two empirical studies are presented to demonstrate this usage. From these studies it is argued that statistical tests of ICC-slopes are not sufficient in testing for heterogeneity. Tests of equality if ICC´s over groups of subject and above all test of equality of ICC´s over groups of items are the proper tests to use. It is also stressed that the application of these kinds of test puts high demands on a theoretical foundation. Finally, it is maintained that the problem of test heterogeneity is essentially a theoretical question.

Table of Contents

1 Introduction.....1
2 The Rasch model and tests of goodness-of-fit.....2
3 Empirical studies.....4
3.1 The Rotating Wheels Tests.....5
3.2 The Spada study.....6
3.2.1 Subjects.....6
3.2.2 Analyses.....6
3.3 The replication study.....14
3.3.1 Subjects.....14
3.3.2 Rasch analyses.....15
3.3.3 Modeling analyses.....17
4 Discussion.....19
References.....22

1 Introduction

An important issue in psychological measurement is the dimensionality of the tests that are used to measure some property of individuals. Several authors have stressed that measurement scales must be homogenous or unidimensional (cf Lord & Novick, 1968; Lumsden, 1976). The meaning of this is that a scale measures one property and not a mixture of several properties. In measurement of intelligence, for example, test should be constructed to measure one ability and not several, i.e. in solving the test items individuals must resort to one ability and not several. The essential concern here is one of interpretability.

Classical test theory is not very well apt to deal with the issue of unidimensionality (cf Lumsden, 1976), even though factor analysis sometimes provides a means to explore test heterogeneity (Levy, 1973). Latent trait theory, however, give us some useful tools. A latent trait model of special interest here is the Rasch model. One reason for this is that this model assumes unidimensionality of observations. A fit of data to the Rasch model can then be regarded as a criterion of unidimensionality.

If a model is to be used as a criterion of for example unidimensionality it is also of vital importance to have the means for detecting violations of the model assumptions. In Rasch model there are within the CML (conditional maximum likelihood) approach such means in the form of statistical tests that are sensitive to different forms of deviations from the model (cf Gustafsson, 1980).

In this paper we demonstrate the use of the Rasch model as a criterion of unidimensionality in of two empirical studies. After that we will discuss some important points of this usage. Before describing the empirical studies we need, however, to take a closer look on the Rasch model and it's goodness-of-fit tests.

2 The Rasch model and tests of goodness-of-fit.

In the Rasch model the probability of a correct answer to an item is a function of two parameters. One parameter describes the difficulty of the item (σ_i, i=1,....,k) and one parameter describes the ability of the tested subject (ξ_v, v=1,....,n). Denoting a correct answer to item i by subject v as A_{vi}=1 we have:

$$P(A_{vi}=1|\xi_v,\sigma_i)=\frac{\exp(\xi_v-\sigma_i)}{1+\exp(\xi_v-\sigma_i)}.$$

There are three basic assumptions of the Rasch model: the assumption of unidimensionality, the assumption of local statistical independence, and the assumption of equal item discrimination. For the present purposes the assumption of unidimensionality is the most important one, even though it has been argued that it is impossible or very difficult to uphold a distinction between the assumptions (Gustafsson, 1980). The meaning of the assumption of unidimensionality is that there is only one latent trait Y affecting performance on all items in the test.

The probability of a correct answer to an item (i) can be expressed as a function of the ability variable (Y). This function is called the item characteristic curve (ICC) and can be expressed in the following way:

$$f_i(\xi)=\frac{\exp(\xi-\sigma_i)}{1+\exp(\xi-\sigma_i)}.$$

In a similar way the probability of a correct answer for each person can be shown as a function of item difficulty to form a person characteristic curve (PCC).

There are several different goodness-of-fit tests that may be used to evaluate the fit of data to the model. The underlying rationale is expressed by Gustafsson (1980):

> "It does seem that the Rasch model can be violated in basically two ways: either a model is needed to describe the data which contains two or more parameters for each person, which would be a violation of the assumption of unidimensionality; or a model is needed which contains two or more parameters for each item, which would be av violation of the assumption of the form of the ICCs; or, of course, a combination of these. If the Rasch model holds true for a set of data the item parameters are invariant from one group of persons to another and the person parameters are invariant from one group of items to another"(ibid, p. 209).

There are two types of test that can be used. One type is called ICC tests, since they are sensitive to violations of the assumption of equal ICC´s. The other type is called PCC tests, since they are sensitive to differences in the PCC´s.

The computer program used to perform the analyses is PML (Gustafsson,1979) which uses conditional maximum likelihood estimates (CML) of the item parameters. This program basically offers three different goodness-of-fit tests and one graphical test of deviations from the model assumptions:

a. The Anderson conditional likelihood ratio test (Anderson,1973). This test essentially is a test of the hypothesis that item parameters are invariant over different groups of persons. When the grouping is done according to level of performance (high-low) on the total set of items this test is called the A-ICCSL test, since it is sensitive to differences in the slopes of the ICC´s. When the grouping is done according to other criteria such as, for example, age or sex, it is called the A-ICC test. With this kind of grouping it is essentially a test of unidimensionality.

b. The Martin-Löf chi-square test (Martin-Löf,1973). This test too is sensitive to heterogenous slopes of the ICC´s and is referred to as the ML-ICCSL test. It is calculated by forming a chi-square sum of the deviations between observed and predicted frequencies of correct answers for each score group. The ML-ICCSL and the A-ICCSL tests are thus sensitive to the same type of deviations.

c. The Martin-Löf test of homogeneity of two sets of items. This is a test of the hypothesis that two sets of items measure the same ability. With items grouped according to level of difficulty it is sensitive to differences of the PCC slopes and, accordingly, it is called the ML-PCCSL test. With items grouped according to other criteria, which should be theoretically derived, it is sensitive to deviations from unidimensionality and is called the ML-PCC test.

3 Empirical studies.

The empirical illustrations are taken from two studies concerned with the analysis of a spatial-mechanical test: "Rotating Wheels". The first study is a reanalysis of a study conducted by Spada (1976) carried out by Lindström (1983a). The second study was carried out by Lindström (1983b) as a partial replication of the Spada study.

There are two reasons for choosing these data. In the first place, Spadas data are well explored (cf Spada, 1976, 1977). In the second place, and more important, the items in the Rotating Wheels Test are fairly well defined and structured.

3.1 The Rotating Wheels Tests.

The test consists of items which depict a system of two or more wheels that are rotating. The subjects are to decide in what direction one or two wheels rotate, given the direction of rotation of another wheel in the system. The system of wheels constituting the items is in principle constructed from four basic components, each including a specific type of tramsmission. In figure 1 below the components are shown.

Insert Figure 1 about here

Component 1 consists of two wheels that are in direct contact with each other. The wheels thus rotate in opposite directions. Component 2 consists of two wheels on a common axis. Obviously, the wheels must rotate in the same direction. Component 3 consists of two wheels that are connected with a transmission band. The wheels then must rotate in the same direction. Component 4 also includes a transmission band, but in this case the band is crossed and the wheels accordingly rotate in opposite directions.

The items are constructed by combining these components in different ways. The structure of the items in the tests used in Spada's study and in the replication study is shown in table 1 below. In the table it can be seen that the first fifteen items are the same in both tests, and that another four items are common, but presented in a different order.

Insert Table 1 about here

In each problem, except the first four, subjects have to decide on the direction of rotation of two wheels. In order to reduce guessing, which obviously must be a problem with only two alternative directions to choose among, there also is a third alternative for the students, namely to say

that they don't know. In figure 2 below an example of one of the items in the tests is shown. All items in the test are constructed according to the same principle.

Insert Figure 2 about here

3.2 The Spada study.

In this section some of the reanalyses of the Spada study is presented. For a full presentation the interested reader is referred to Lindström (1983a).

3.2.1 Subjects.

There were 518 subjects in the sample. The subjects were of age 11 to 16 attending German secondary school. There were 310 pupils from grades five and six (age 10 to 13) and 208 pupils from grades eight and nine (age 13 to 18). Both sexes were represented, with 253 female and 265 male subjects.

3.2.2 Analyses.

In the following presentation we first give descriptive results of the test. After that results of the ICCSL-tests will be presented, followed by results of the PCC- and PCCSL-tests.

In table 2 below the proportion of correct answers on the items and the mean and standard deviation of the whole test are shown. The proportions of correct answers generally are quite high, which also is reflected in the mean of 14.30 on 20 items. The highest value is .97 and the lowest is .56. From the results on the first four "single component" items it can be seen that components 2 and 3 are fairly easy and that component 1 is the most difficult one.

Insert Table 2 about here

The reliability of the whole test is, accordning to the KR-20 formula, .88, which is rather good for a test of 20 items. To a large extent, however, the high reliability reflects the fact that the sample is quite heterogenous with respect to level of ability.

In testing the fit of data to the model the ML-ICCSL test and the A-ICCSL test were performed in the first step. Both these tests are sensitive to differences in the slopes of the ICC's. The ML-ICCSL test gave a chi-square value of 398.8 with 304 degrees of freedom (p=.0002) and the A-ICCSL test gave a chi-square value of 56.3 with 19 degrees of freedom (p=.00001). Both tests thus show a significant deviation of data from the model. A closer inspection of the results show that the biserial correlations for the four introductory items and item 19 are comparatively lower, which is an indication that these items do devitate from the rest of the items. This may also be the case for item 5. Furthermore, a t-test of the differences between item parameters estimated for low scoring subjects (with a raw score between 1 and 11) and high scoring subjects (with a raw score between 12 and 19) shows a highly significant difference for item 19. This item is relatively easy for the low subjects and relatively difficult for the high subjects. Spada (1976, p 219) explained this by motivational factors. Subjects with high scores were not motivated to invest energy in trying to solve items of this complexity towards the end of the test, since they might feel they know the problem and the work is felt to be monotonous.

There are reasons to reject item 19 from further analyses. It is also reasonable to discard the four introductory single-component items. The reason is that these were accompanied by instructions in order to make subjects familiar with the type of problems. Item 5 will be kept since there are no strong substantive reasons to reject it. The fact that it is the first item without instruction and the first item asking for the rotation of two wheels is not sufficient.

We are thus left with a set of fifteen items and the fit of these data are subsequently tested with the use of the ICCSL tests. The ML-ICCSL test gave a chi-square value of 198.9 with 182 degrees of freedom (p=.19) and the A-ICCSL test gave a chi-square value of 16.0 with 14 degrees freedom (p=.31), which indicates a good fit of data to the model.

The analyses will be continued by testing if the estimates of the item parameters are equal in subgroups when these subgroups are defined from external criteria, thus performing an A-ICC test. The reason for this is that there may be violations of model assumptions that are not detected by the ICCSL tests since they are sensitive only to differences in the slopes of the ICC's. It would be unsatisfactory to accept a fit of data to the model on the basis of the nonsignificant results of the ICCSL tests since multidimensionality would not be detected with these tests as long as it is not reflected in different item discriminations.

One criterion of interest for this kind of test is sex and another is age. Sex is interesting since it may be hypothesized that boys and girls differ in the way they solve problems of spatial nature, girls to a larger extent relying on reasoning (cf French, 1965; Gustafsson, 1976; Gustafsson, Lindström & Björck-Åkesson, 1981; Lohman, 1979; Snow, 1976; Vandenberg, 1975). If in fact boys and girls differ in this respect it would be possible to detect multidimensionality with the A-ICC test, at least to the extent that the strategies can be applied with a different degree of success on different items. The same argument is valid for subjects of different age. Older subjects may to a larger extent try to develop strategies to handle the problems. They may also have real life experiences of rotation problems, and especially so boys, that form a basis for the development of such strategies.

Boys are generally better at solving rotation problems than girls. The mean for the boys is 11.39 and the standard deviation 3.71. Girls have a mean of 8.93 with a standard deviation of 4.28. The A-ICC test gave a chi-square value of 13.8 with 14 degrees of freedom (p=.47). This result does not indicate that the test is multidimensional.

The younger subjects (grades 5 and 6) have a mean value of 9.14 and a standard deviation of 4.22. Older subjects (grades 8 and 9) have a mean value of 11.75 and a standard deviation of 3.60. The smaller standard deviations for older subjects and for boys (see above) is evidently due to a ceiling effect. The A-ICC test gave a chi-square value of 16.6 with 13 degrees of freedom (p=.22). To judge from the ICC tests, there are no reasons to reject the model.

However, it would be premature to accept a fit of data without considering internal criteria. That is, we must consider differences between individual items or groups of items on substantive grounds. In this case the obvious thing to do is to examine the different components to see whether they may differ in what they demand of the subjects. We can then group the items according to hypothesized differences and with this grouping as a basis we can perform an ML-PCC test.

Lindström (1983a) discussed possible differences between the components. Here only a shorthand version of the arguments will be given. All the components to a smaller or larger extent give possibilities to rely on different abilities to arrive at a solution. Subjects may use reasoning processes and they may use visualization processes, or some combindaation of them. Making the case that the Rotating Wheels test is a test of spatial visualization we may examine the possible differences between the components.

From an analysis of the correlational literature on spatial ability Lohman (1979) concluded:

> "the crucial component of spatial thinking may be the ability to generate a mental image, perform various transformations on it, and remember the changes in the image as the transformations are performed. This ability to update the image may imply resistance

to interference, both externally and internaly generated. Further, it implies that one of the crucial features of individual differences in spatial ability may lie not in the vividness of the image, but in the control the imager can exercise over the image (Lohman, 1979, p. 116).

Individual differences in spatial ability may be explained by differences in the ability to perform processes of the kind outlined, but there are also other explanations. Gustafsson et al (1981) argued that another line of reasoning is to take the starting point in differences in long-term memory as a consequence of prior learning in the figural/spatial domain. These explanations are by no means mutually exclusive, they are rather complementary. The relationships between the different mechanisms may however be rather intricate. Thus, small initial differences in the ability to perform certain processes may influence interests and preferences which in turn may cause large differences in aquired knowledge. On the other hand, experiences in a domain may also influence the ability to perform processes of a specific kind (cf Salomon, 1979).

Starting from these generel considerations of spatial abilities differences between the components can be discussed in terms of two aspects; the possibilities of holistic processing, and the demands the components put on subjects to find a "starting point" for visualizing.

The four components differ with regard to the possibilities for holistic processing. In components 2 and 3 both wheels rotate in the same direction and the whole construction form a global unit which is rather easy to visualize. (It is in fact also a simple logical derivation that the wheels should rotate in the same direction.) Components 1 and 4 differ in this respect. Especially in component 1 there is a direct contradiction between the movements of the wheels and the formation of a global unit based on the holistic characteristics of the perceptual field.

The other aspect, to find a starting point for visualizing the movements of the wheels is not essential for components 2 and 3. This starting point is in fact directly given by the perceptual features of the visual field. In component 4 this starting point is externally provided by the twisted transmission belt. There is then a slight difference between component 4 and component 2 and 3 in this respect. Component 1, however, differes more

radically from the rest of the components since subjects themselves have to provide a "mediator". This can be phrased in terms of prior knowledge of for example rolling machines.

The problem of finding a starting point can also be phrased in terms of a necessity of focusing on a part of the visual field or problem, from a structuring of it.

This way of discussing the different components essentially conforms to a hierarchical model of cognitive processing, where there are higher order processes directing the utilization of lower ones. This applies both to the field of differential psychology (cf Gustafsson, Lindström, & Björck-Åkesson, 1981) and cognitive/learning psychology (Lindström, 1983c).

The discussion above implies that component 1 and component 4, on one hand, and components 2 and 3, on the other, are different with regard to what abilities they may call for. Also components 1 and 4 are different from each other. This should then be reflected in the items.

It is possible to test the hypothesis of equal PCC slopes, that is, homogeneity, with the ML-PCC test. In this test the items have to be divided into two groups and the only way of doing this is to have one set of items including a certain component and one set of items not including the component in question. In table 3 below the results from these tests are summarized.

Insert Table 3 about here

Apart from the comparisons discussed above, a comparison of items including component 1 but not component 4 with items including component 4 but not component 1 is done. It should then be noted that in this analysis there are only three items in each subset.

The results give evidence for the hypothesis that the test is heterogenous

and that one main source of this is differences between the components.

To take a closer look at the differences between the components the analyses can be carried further by using data on how subjects solved some of the single components within the more complex tasks. These analyses are fully presented in Lindström (1983a) and here only a summary of the results are given.

In order to answer the more complex items in the test sujects have to give the direction of rotation of two of the wheels in a problem. In order to be correct on an item subjects must be correct about both answers. This makes it possible to see if a subject is correct about one or two of the single components in a task, naturally depending on which specific wheel apart from the last one is in focus. Subjects may be wrong about the direction of rotation of the first wheel and be right about the direction of rotation of the second wheel, given the wrong direction of the first wheel. Data were available that made it possible to make this kinds of judgements, and thus decide upon the correctness of single components in a task.

When analyzing this set of components only one component from each item has been included in the total set of components. The single component items 1 to 4 are also excluded, for reasons mentioned above. The reason why only one component from each item is used is that two components from the same item would violate the assumption of local statistical independence, that is, the probability of answering both components correctly is the product of the probabilities of answering each component correctly. This assumption is violated by the fact that the components have the same context and probably this context will have a strong influence on the problem solving, and influence answers on both components in the same way.

When the fit of a set of eleven item components was tested with the ML-ICCSL test a chi-square value of 151.1 with 90 degrees of freedom ($p=.0006$) was found. The A-ICCSL test gave a lower p value ($p=.016$) but this test is less sensitive to other deviations in the data than different discrimination, with

the automatic grouping used. There was indications of one of the component items (C2-5) showing a poor fit. Rejecting this item and retesting the scale of ten item components by the ML-ICCSL test gave a higher, but still significant, p value (p=.006). The reliability of the scale of eleven item components was .69.

A ML-PCC test, with component 1 in one group and the other item components in the other group gave a highly significant chi-square value of 87.7 with 23 degrees of freedom (p=.000005). There are only three items remaining in the component 1 group. An ML-ICCSL test of how these item fit together showed a non significant chi-square value (p=.81). An ML-PCC test of item components 2 and 3 versus item component 4 gave a chi-square value of 13.2 with 14 degrees of freedom (p=.51). In this test there are no indications that these item components differ.

It is evident from the results of the analyses presented above that the Rotation Wheels test is a heterogenous test. This heterogeneity does not appear to follow a simple model which means that the test can not be divided into several homogeneous subscales. Instead, the items in the test seem to be inherently heterogenous The reason for this is that the items are constructed of components making demands on different abilities.

The question of guessing is pertinent with respect to the construction of this test. It might have been expected that subjects would guess to a rather high extent, since there are only two alternatives to choose between in each step. However, there are no indications of guessing in the data and it appears that the subjects have followed the instructions not to guess.

3.3 The replication study.

The replication study was carried out as a replication of the Spada study with a Swedish sample of subjects. The original Rotating Wheels test was revised on the basis of the reanalysis of Spada's data. Nineteen out of the twenty items in the original test were kept and the test was extended by adding items of the same basic construction (see table 1). The revised test consisted of a total of 35 items. The format of the test was furthermore changed, mainly by making the drawings of the constructions of wheels more "realistic". This was done to possibly enhance visualization.

In order to study the validity of the Rotating Wheels test it was also judged important to be able to relate the results on the test to other psychometric data, notably on spatial abilities. This was obtained by drawing the sample for this study from a "reference population", for which already psychometric data on a wide range of abilities was available.

In this paper only a part of the the analyses carried out will be presented. We will restrict ourself to presenting Rasch analyses of the part of the test that is the same as in the original test. This will be sufficient for the present purposes. In the presentation of some modeling analyses we will however use more of the test items.

3.3.1 Subjects

There were 341 subjects in the sample, 181 boys and 160 girls. These subjects were drawn from a reference population of 1254 pupils in grade 6 (around 13 years old) in the Swedish compulsory school.

3.3.2 Rasch analyses

The analyses presented here follow the same pattern as the analyses of the Spada data. The substantive reasons given for performing the specific analyses are given in the description of the analyses of the Spada study and will not be repeated.

In table 4 below the proportions of correct answers to the items are presented alogn with summary statistics. In addition to the results given in the table it should be mentioned that the reliability for the subset of items 5 to 15 was .86 (KR-20), which is rather good.

Insert Table 4 about here

When items 5 to 15 were analyzed the ML-ICCSL test gave a non-significant chi-square value of 109.1 (df=90, p=.08). The A-ICCSL test on the other hand gave a significant chi-square value of 23.3 with 10 degrees-of-freedom (p=.001). Examination of possible reasons for the deviations from the model revealed two sources.

In the ML-ICCSL test it is shown that the main contribution to the total chi-square value come from the low scoring groups. The group scoring a total of 1 point made a contribution of 24 percent, the group scoring a total of 2 made a contribution of 15 percent, and the group scoring 3 made a contribution of 10 percent. It is however difficult to draw any conclusions from this observations, especially since the number of subjects in the lowest scoring group is quite low (n=12).

There are systematic differences between the estimates of item parameters for the low scoring and high scoring groups respectively. Items without component 1 are systematically more difficult in the low scoring group, while items including component 1 are more difficult in the high scoring group. (Only one of the differences is actually significant at the five-percent level, but the

pattern is completely systematic.)

When items 16-28 were analyzed the ML-ICCSL test gave a chi-square value of 140.5 with 132 degrees of freedom (p=.29) and the A-ICCSL test gave a chi-square value of 16.9 sith 12 degrees of freedom (p=.15). A closer examination of the results indicated that also in this set of items there is a tendency of systematic differences in estimated difficulties of the items even though the pattern is not as systematic as among items 5-15.

Also in these data a A-ICC test was carried out. Sex was used as an external criterion, for reasons given above. The test gave a chi-square value of 7.6 with 10 degrees of freedom (p=.67). There are thus no indications indications of differences between the sexes.

The analyses of these data were carried out as a replication of the analyses of the Spada study. A set of PCC test was carried out to test differences between the components and the results of these analyses are summarized in table 5 below.

Insert Table 5 about here

The conclusion that can be drawn from these analyses is that there are strong indications that component 1 differ from the other components. It can be added that the same conclusion is drawn from an analyses of the rest of the items in the test (Lindström, 1983b). This set essentially behave in the same way as the replication set, apart from the fact that the ICCSL test did not indicate any deviations from the model. The pattern of results of the ML-PCC tests was however exactly the same. Analyses of component data have also been carried out. These analyses confirmed the conclusions drawn from the analyses of item data presented here.

3.3.3 Modeling analyses.

The relation of the Rotating Wheels test to other test performances was studied within the LISREL framework (Jöreskog, 1974) using the LISREL V program (Jöreskog & Sörbom, 1981). The LISREL-model used here consists of two parts, the measurement models for the psykometric test data and for the Rotating Wheels test, respectively. The measurement models are factor analytic models. Structural relations are then specified between the two measurement models. In this case these relations correspond to regression coefficients. In summary the LISREL-model used here can be described as a regression model with latent variables, where Rotating Wheels is regressed on background variables.

In this paper only one of the models actually tested is presented. The different steps taken in arriving at this model are neither presented. The interested reader is referred to Lindström (1983b), which discuss these issues and also involves a more full account of the sustantive issues. (For a discussion of the methodology adopted in the LISREL-analyses see Gustafsson & Lindström, 1979; and Gustafsson, Lindström & Björck-Åkesson, 1981.)

The measurement model used for the psykometric test data is a general model for the organization of cognitive abilities, the HILI-model (hierarchical lisrel based model, Gustafsson, Lindström & Björck-Åkesson, 1981). At the primary level there are seven factors Vz (Visualization), S (Spatial Orientation), Cf (Flexibility of Closure), CFR (Cognition of Figural Relations), I (Induction), Ms (Memory Span), and V (Verbal Comprehension). At the second-order level there are three factors; Gc (crystallized intelligence), Gv (general visualization), and Gf (fluid intelligence), (Cattell & Horn, 1966). At the third-order level a factor of general intelligence, corresponding to Spearmans g is specified.

The measurement model for the Rotating Wheels test presented here was defined with component data as observed variables. The primary reason for this was to be able to study specific component effects. The sum of correct answers on components was computed within two sets of items (items 5-15 and 16-18,

respectively). There are thus eight observed variables which define two first-order factors, one for set 1 (Rw1) and one for set 2 (Rw2). On the second-order level the model contains one latent variable (Rw). This construction of the model with two first-order factors is done for both methodological and substantive reasons. Methodologically it makes it possible to relate this analysis to the analyses of the Spada data. There are no strong substantive theoretical reasons for dividing the items in two sets as was done. However, this division makes it possible to have a brief look on what might be called "learning effects". Lindström (1983b) presents a closer analysis of these problems in the present data.

In figure 3 below the full model is presented. Looking at the measurement model for the Rotating wheels test it can be seen that the latent variable RW1 is defined by the component data from set 1 and the variable RW2 is defined by the four observed variables in set 2. It can also be seen that component 1 answers in set 2 have a small loading in RW1. In LISREL it is also possible to specify correlations between the residuals and in the model there are correlations between the residuals for components 1, 2 and 4 over the two sets. This also indicates the presence of component specific factors. Component 3 is an exception in this case. The correlation between the residuals for the component 3 variables approached zero and was naturally not significant. In the final model this correlation was therefore specified to be zero.

Insert Figure 3 about here

When the measurement model for Rotating Wheels is regressed on the HILI model we find that the dependent second-order factor RW is predicted by Gv and the dependent primary factor RW1 is predicted by Vz. This model had a chi-square value of 252.9 with 210 degrees of freedom (p=.023), which is not too bad a fit, considering the complexity of the model.

In the course of the analyses it was examined if there were any other relations. These analyses gave only small nonsignificant coefficients for G

and Gc on Rw and Vz on Rwl and in the final model these relations were specified to be zero. There are then only direct effects of Gv and Vz. The G factor has, however, an indirect effect through Gv and Gv-Vz.

The model presented here give support for the hypothesis that there are demands for component specific abilities in the test. The correlations between the residuals in the LISREL model indicate more of component specificity than was revealed in the Rasch analyses, but this is of course due to the fact that component data is used in the modelling analyses.

There are, however, other aspects of the model that are more important. In the first place there seem to be some differences between the two sets of data, indicated by the differences in factorial structure and the relation between Vz and Rwl. These differences may reflect "learning effects", which in this case mean that subjects learn the components when working through the test. In later parts of the test subjects need not to solve the single component problems, but may resort to a knowledge of how one wheel effect another once the component is identified. The relation between Vz and Rwl is specifically a support for the idea that in the earlier parts of the test the ability to perform specific visualization processes is more important.

Secondly, and still more important the model suggests that more general abilities play an important role for the performance on the test.

4 Discussion.

The empirical analyses presented above demonstrate a methodology for exploring tests where the Rasch model is used as a criterion of unidimensionality. A number of issues stand forward in such an application of the Rasch model.

In the first place the means of detecting lack of fit of data to the model and violations of specific model assumptions is crucial. This is mainly a

question of what tests there are at disposal and the sensitivity of these tests to different kinds of violations. Gustafsson (1980) argued that the tests appropriate for direct test of unidimensionality are the ICC tests and above all the PCC test. The ICCSL tests are not sensitive to violations of the assumption of unidimensionality, at least not as long as multidimensionality is not reflected in differences in discriminative power. This is, I think, amply demonstrated in the empirical analyses above. The results from the ICCSL-test would have led us to accept a fit of data to the model. The criterion of unidimensionality would have been met and the Rotating Wheels test would have been accepted as a homogenous test. The results of the PCC test show, however, that such an acceptance is premature. These tests clearly show that the Rotating Wheels test in fact is a heterogenous test. The ICC test gave no such indications which of course may depend on the hypotheses underlying the grouping of subjects may be wrong. The main conclusion to draw from this is that the PCC test should have a central place in applications of the Rasch model as a criterion.

Hovewer, both the ICC tests and the PCC test put the necessity of a theoretical starting point into focus. Application of these tests is essentially a question of testing hypotheses about multidimensionality, either based on external considerations (the ICC tests) or internal considerations (the PCC test). The external considerations imply that we are able to formulate hypotheses about heterogeneity of individuals. When internal considerations is concerned the major sources of information comes from a close scrutiny of items and possibly observations of problem solving processes (cf Cronbach, 1970). From such information it is possible to formulate hypotheses about differences between items. This also means that our tests are no better then the hypotheses we can formulate. This is especially crucial in relation to test development, which to a large extent is an atheoretical enterprise, i.e. there is a weak relation between psychological theory and test theory (cf. Levy, 1973). Levy also emphasized the need for test of the same principle conctruction as the ICC and PCC tests saying that: "Much needed are tests constructed to tests hypotheses, and fewer hypotheses about tests (Levy, 1973, p 37).

It is possible to distinguish between different forms of heterogeneity. Levy

(1973) distinguished between three types of models: the compensatory model, the conjunctive model and the disjunctive model. The compensatory or additive model implies the subject low on one ability may compensate that by beeing higher on another ability. The factor analytic model is of this kind. An IQ test may follow such a model. The conjunctive model implies that subjects must be high enough on several abilities to be able to solve a test item. A test item may, for example, demand a certain amount of reasoning ability and a certain amount of spatial visualization ability. The disjunctive model implies that subjects may resort to one of several abilities in order to solve an item correctly. An example of this was given in the empirical analyses presented above. In applying the ICC tests we built on previous research which has hypothesized sex differences with regard to what kind of processes subjects resort to in solving spatial test items (see above, page 8).

Basically a model of multidimensionality following an additive or a disjunctive model would be excluded in the Rasch model. The conjunctive case, however, is a bit more problematic. To the extent that the relation between the dimensions differ between items and this is reflected different rank orderings over items the model would be excluded, otherwise not. On the other hand, if the relation holds over items one could on substantive grounds challange the whole idea of multidimensionality in this case.

Now, what is the value of using the Rasch model as a criterion in theoretically oriented work? I think, first of all, that it is a valuable tool in test construction and test development, since it enables us to explore the properties of the items in the test. It may also be of theoretical interest in itself to perform such analyses. This falls back to the necessity to formulate hypotheses on theoretical substantive grounds. Secondly, I will argue that it can be no objective in itself to have a test or set of items that fit the Rasch model. The ultimate criterion of measurement is what psychological construct we want to measure. In the Rasch analyses we may capture much of what is specific in a test, but there is a risk of not seeing the wood for all the threes. This is to say that a set of items that are multidimensional according to the Rasch criterion could still be a good measure of more general human abilities. In the modelling analysis presented above we tried to illustrate that point.

References

Andersen, E.B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140.

Cronbach, L.J. Test validation. In: R.L. Thorndike (Ed.) Educational Measurement. Washington: American Council on Education, 2nd ed., pp xxxx, 1970.

French, J.W. The relationship of problem-solving styles to the factor composition of tests. Educational and Psychological Measurement, 1965, 25, 9-28.

Gustafsson, J.-E. Verbal and figural aptitudes in relation to instructional methods. Göteborg: Acta Universitatis Gothoburgensis, 1976.

Gustafsson, J.-E. PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items. Reports from the Institute of Education, University of Göteborg, no 85, 1979.

Gustafsson, J.-E. Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 1980, 33, 205-233.

Gustafsson, J.-E., & Lindström, B. Analyzing ATI data by structural analysis of covariance matrices. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Gustafsson, J.-E., Lindström, B., & Björck-Åkesson, E. A general model for the organization of cognitive abilities. Reports from the Department of Education, University of Göteborg, 1981:06.

Horn, J.L., & Cattell, R.B. Refinement and test of the theory of fluid and crystallized intelligence. Journal of Educational Psychology, 1966, 57, 253-270.

Jöreskog, K.G. Analyzing psychological data by structural analysis of covariance matrices. In: R.C. Atkinson, D.H. Krantz, & R.D. Suppes (Eds.) Contemporary developments in mathematical psychology, Volume II. San Francisco: W.H. Freeman., pp 1-56, 1974.

Jöreskog, K.G., & Sörbom, D. LISREL V. Analysis of linear structural relationships by maximum likelihood and least squares methods. Research report 81-8, University of Uppsala, Department of Statistics, 1981.

Levy, P. On the relation between test theory and psychology. In: P. Kline (Ed.) New approaches in psychological measurement. London: John Wiley., pp 1-42, 1973.

Lindström, B. Rotating wheels - An analysis of a test of spatial ability using the Rasch model. Reports from the Department of Education, University of Göteborg, 1983:02 (a).

Lindström, B. Rotating Wheels - A replication study of a test of spatial ability. Department of Education, University of Göteborg, 1983 (b).

Lindström, B. Learning styles and learning strategies. I. Conversation theory - The work of Gordon Pask. Reports from the Department of Education, University of Göteborg, 1983:01 (c).

Lohman, D. Spatial ability: A review and reanalysis of the correlational literature. Technical report no. 8, Aptitude Research Project, School of Education, Stanford University, 1979.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading: Addison-Wesley, 1968.

Lumsden, J. Test theory. In: M.R. Rosenzweig, & L.W. Porter (Eds.) Annual Review of Psychology, 27. Palo Alto: Annual Reviews Inc., pp 251-280, 1976.

Martin-Löf, P. Statistiska modeller. Anteckningar från seminarier läsåret 1969-70 utarbetade av Rolf Sundberg. 2:a uppl. (Statistical models. Notes from seminars 1969-70 by Rolf Sundberg. 2nd edition). Institutet för försäkringsmatematik och matematisk statistik vid Stockholms universitet, 1973.

Salomon, G. Shape, not only content: How media symbols partake in the development of abilities. In: E. Wartella (Ed.) Children communicating: Media and development of thought, speach, understanding. London: Sage Publications., pp XXXX, 1979.

Snow, R.E. Learning and individual differences. In: L.S. Shulman (Ed.) Review of research in education Volume 4. Itasca: F.E. Peacock Publishers, Inc., pp XXX, 1976.

Spada, H. Modelle des Denkens und Lernens. Bern: Verlag Hans Huber, 1976.

Spada, H. Logistic models of learning and thought. In: H.Spada, & W.F. Kempf (Eds.) Structural models of thinking and learning. Bern: Hans Huber Publishers., pp XXX, 1977.

Vandenberg, S.G. Sources of variance in performance of spatial tests. In: Eliot, J., & Salkind, N.J. (Eds.) Children's spatial development. Springfield, Illinois: Charles C Thomas Publisher, pp 57-66, 1975.

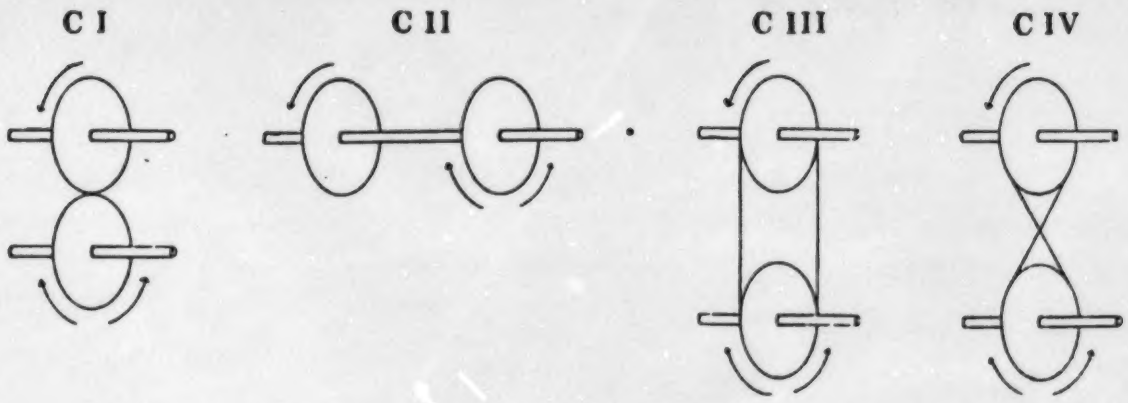

Figure 1. The components in the rotating wheels problems.

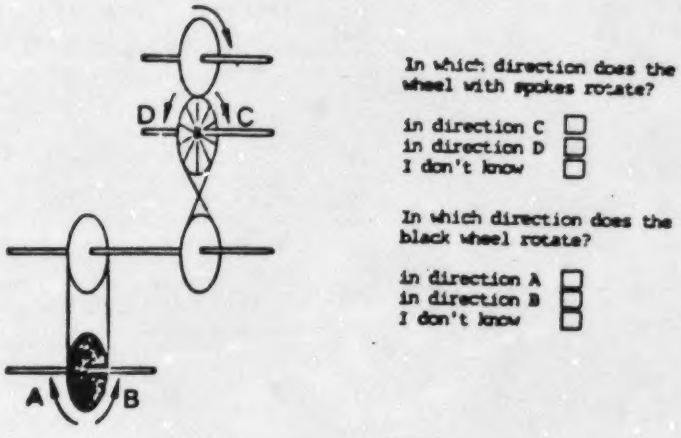

Figure 2. An example of an item in the Rotating Wheels test (Item 10 in Spada´s and Lindström´s tests.)

BEST COPY AVAILABLE

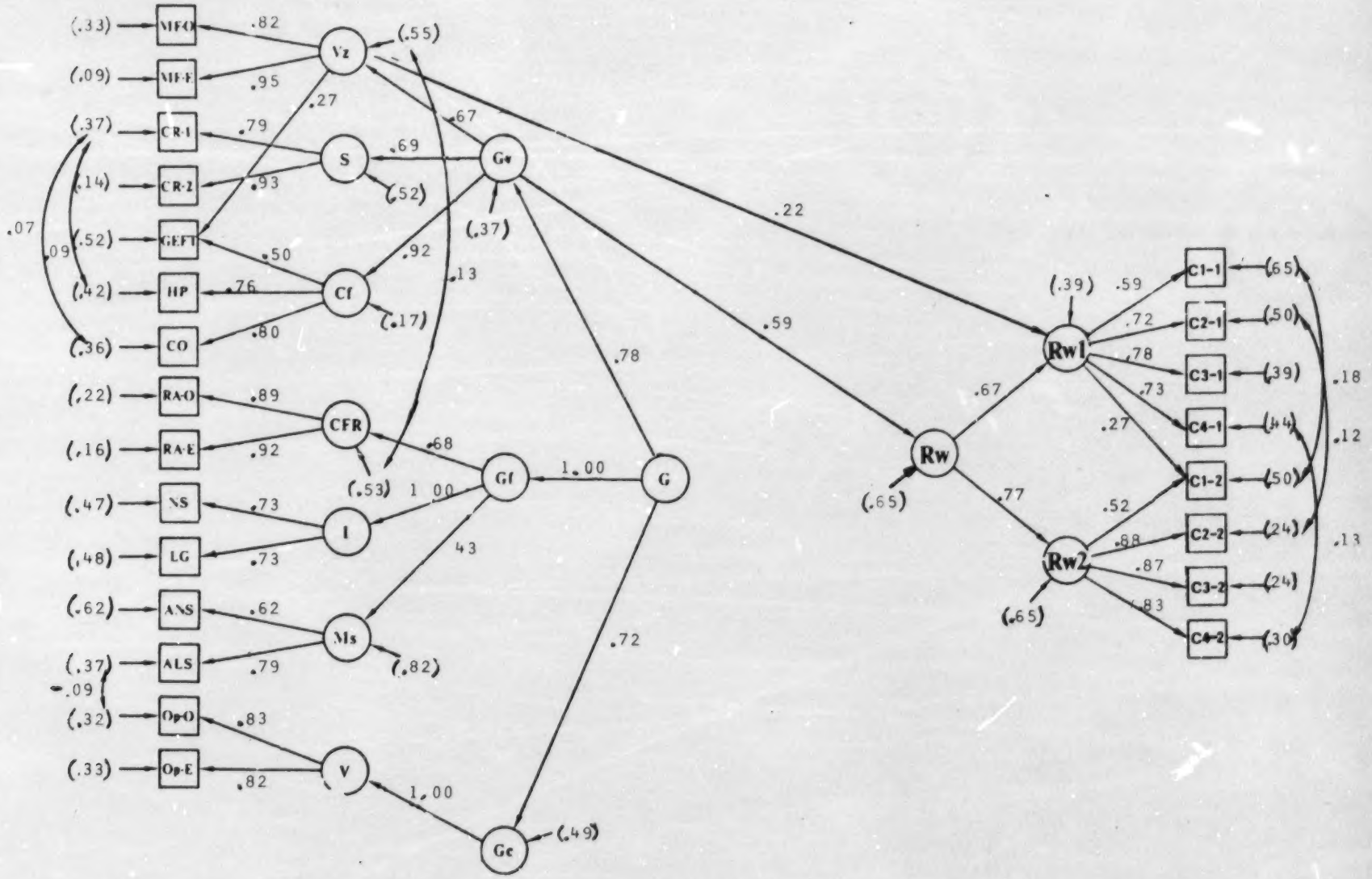

Figure 1. A LISREL model for the relations between the HILI model and a model of the Rotating Wheels test.

Table 1. Task structure of the rotation problems in Lindström's study (RW-L) and in Spada's study (RW-S).

Item no RW-L	Item no RW-S	Component I	II	III	IV	Number of components
1	1	1	0	0	0	1
2	2	0	1	0	0	1
3	3	0	0	0	1	1
4	4	0	0	0	1	1
5	5	1	1	0	0	2
6	6	1	0	1	1	3
7	7	1	0	0	1	2
8	8	1	1	1	0	3
9	9	1	0	1	0	2
10	10	1	1	1	1	4
11	11	0	0	1	1	2
12	12	1	1	0	1	3
13	13	0	1	0	1	2
14	14	0	1	1	1	3
15	15	0	1	1	0	2
16		1	1	0	0	2
17		2	0	0	1	3
18		2	0	0	0	2
19		1	0	0	1	2
20		2	1	0	0	3
21		0	0	3	0	3
22		1	0	1	0	2
23		0	1	0	1	2
24		2	0	1	0	3
25		0	1	1	0	2
26		0	2	0	0	2
27		0	0	1	1	2
28		0	0	0	2	2
29	16	1	1	1	2	5
30		1	1	1	2	5
31	17	1	2	1	1	5
32		2	2	1	1	6
33	18	2	1	1	1	5
34		2	1	2	1	6
35	20	1	1	2	1	5
	19	2	2	2	2	8

Table 2. Proportions of correct answers on the items together with mean and standard deviation of the Rotating Wheels test in the Spada (1976) study.

Item no.	Proportion of correct answers.	Item no.	Proportion of correct answers.
1	.74	11	.81
2	.93	12	.66
3	.97	13	.84
4	.87	14	.83
5	.60	15	.89
6	.60	16	.59
7	.57	17	.59
8	.65	18	.56
9	.71	19	.61
10	.65	20	.61

Mean = 14.30, S.D. = 23.38, Reliability(KR-20) = .88

Table 3. ML-PCC tests of differences between subsets of items in Spada's (1976) study.

Contrast*	Chi-square	df	p
Items with C1 vs items without C1	109.92	43	.00000
Items with C4 vs items without C4	102.88	43	.00000
Items with C2 vs items without C2	54.84	43	.106
Items with C3 vs items without C3	53.86	43	.124
Items without C4 and with C1 vs Items without C1 and with C4	80.49	8	.00000

* For information about the specific items included in each subset reference is made to table 1.

Table 4. Proportions of correct answers on the items together with mean and standard deviation of the Rotating Wheels test in Lindström's replication study.

Item no.	Proportion of correct answers.	Item no.	Proportion of correct answers.
1	.86	16	.69
2	.87	17	.58
3	.92	18	.63
4	.79	19	.60
		20	.56
5	.60	21	.67
6	.50	22	.65
7	.62	23	.63
8	.61	24	.56
9	.62	25	.63
10	.55	26	.64
11	.64	27	.53
12	.57	28	.54
13	.73		
14	.67	29	.36
15	.77	30	.35
		31	.32
		32	.28
		33	.33
		34	.28
		35	.25

Mean = 20.39, S.D. = 8.81, Reliability (KR-20) = .93

Table 5. ML-PCC tests of differences between subsets of items in Lindström's replication study.

Contrast*	Chi-square	df	p
Items with C1 vs items without C1	70.62	27	.00001
Items with C2 vs items without C2	28.66	27	.38
Items with C3 vs items without C3	27.83	27	.41
Items with C4 vs items without C4	32.80	27	.20

* For information about the specific items included in each subset reference is made to table 1.